

ADAM POUR LE DEEP LEARNING

ACHIQ Aya, CLETZ Laura, EL MAZZOUJI Wahel

Octobre 2025



UNIVERSITÉ DE
MONTPELLIER



STATISTIQUE
SCIENCE DES DONNÉES
UNIVERSITÉ DE MONTPELLIER



FACULTÉ DES SCIENCES
DE MONTPELLIER

SOMMAIRE

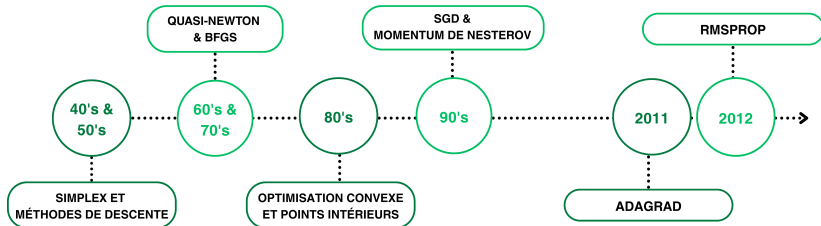
ENJEUX ET CADRE STATISTIQUE

FORCES ET FAIBLESSES

CONCLUSION

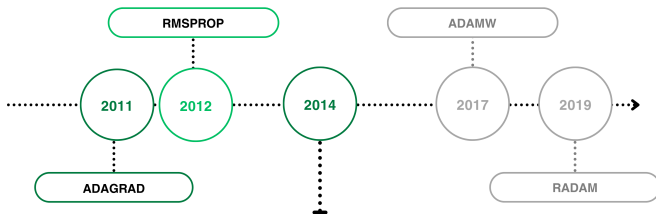
OPTIMISATION EN DEEP LEARNING

- ▶ But : minimiser le fonction de **perte** $\mathcal{L}(\theta)$ pour un poids θ .
- ▶ Enjeux :
 - ▶ Convergence rapide ;
 - ▶ Stabilité numérique ;
 - ▶ Bonne généralisation.



OPTIMISATION EN DEEP LEARNING

- ▶ But : minimiser la fonction de **perte** $\mathcal{L}(\theta)$ pour un poids θ .
- ▶ Enjeux :
 - ▶ Convergence rapide ;
 - ▶ Stabilité numérique ;
 - ▶ Bonne généralisation.



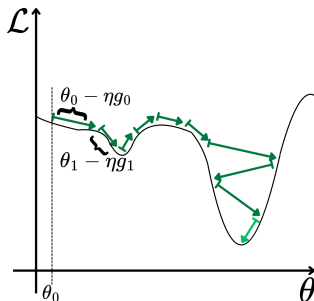
ADAM, KINGMA et BA, 2014

- ▶ Est-ce l'algorithme d'optimisation idéal pour le Deep Learning ?

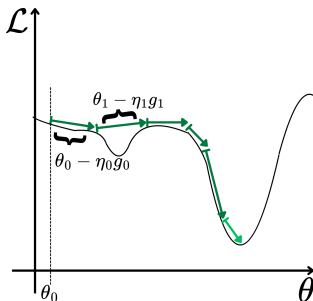
NOTIONS D'OPTIMISATION

► Notations :

- Le gradient $g_t = \nabla_{\theta} \mathcal{L}(\theta_{t-1})$;
- Le **learning rate** η : contrôle la taille des **pas** de mise à jour des paramètres, diffère suivant la méthode (RUDER, 2016) ;
- Le **momentum** m_t : lissage de la trajectoire des gradients, calculé à partir de g_{t-1} .



Stochastic Gradient Descent



Root Mean Square Propagation

PRINCIPE DE L'ALGORITHME ADAM

IDÉE CLÉ

Adam = Adaptive Moment Estimation : combine le **momentum** et une adaptation du pas pour chaque paramètre.

- Moyennes mobiles :

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t, \quad v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$$

- Mise à jour :

$$\theta_t = \theta_{t-1} - \alpha \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \varepsilon}$$

- Valeurs typiques : $\alpha=0.001$, $\beta_1=0.9$, $\beta_2=0.999$.

Propriétés : invariance d'échelle, stabilité, faible sensibilité aux hyperparamètres.

OBJECTIFS ET LIMITES D'ADAM

Objectifs :

- ▶ Accélérer la convergence avec momentum et adaptation des pas ;
- ▶ Stabiliser l'apprentissage en présence de gradients bruités ;
- ▶ Adapter automatiquement le taux d'apprentissage pour chaque paramètre.

Limites :

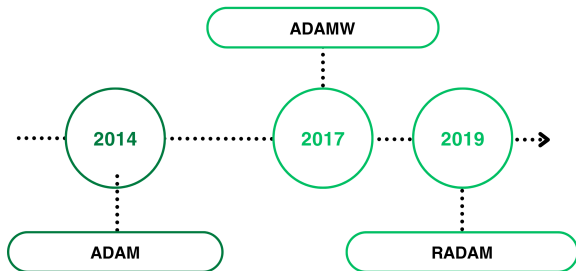
- ▶ Moindre généralisation que SGD WILSON et al., 2017 ;
- ▶ Convergence vers $\|w\|_\infty$ minimale ;
- ▶ Compromis entre vitesse et généralisation.

EXPÉRIMENTATION

EXPÉRIMENTATION

CONCLUSION

- ▶ Avantages d'Adam :
 - ▶ rapidité à l'initialisation ;
 - ▶ convergence rapide par momentum ;
 - ▶ stabilité numérique en présence de gradients bruités ;
 - ▶ adaptation automatique des pas.
- ▶ Limites d'Adam :
 - ▶ faible capacité de généralisation (surapprentissage) ;
 - ▶ dépendance de la qualité des données.



RÉFÉRENCES

Kingma, Diederik P. et Jimmy Lei Ba (2014). « Adam : A Method for Stochastic Optimization ». In : url :

<https://arxiv.org/abs/1412.6980>.

Ruder, Sebastian (2016). « An overview of gradient descent optimization algorithms ». In : url :

<https://arxiv.org/pdf/1609.04747>.

Wilson, Ashia C. et al. (2017). « The Marginal Value of Adaptive Gradient Methods in Machine Learning ». In : url :

<https://arxiv.org/abs/1705.08292>.